Etude de la Maximisation de l'Influence dans les Réseaux Sociaux

Cédric Lagnier, Gaussier Eric

Etude de la Maximisation de l'Influence dans les Réseaux Sociaux

Cédric Lagnier — Eric Gaussier

Université Grenoble 1, LIG
UFR IM2AG - BP 53 - 38041 Grenoble Cedex 9 - France
{prenom.nom@imag.fr}

RÉSUMÉ. *La maximisation de l'influence est un problème NP-difficile lié à la diffusion de l'information dans les réseaux sociaux. L'algorithme glouton "Greedy hill climbing" en fournit une bonne approximation lorsque la fonction d'influence à optimiser est sous-modulaire, ce qui est le cas pour les modèles de diffusion standards. Nous présentons ici un modèle de diffusion non équivalent aux modèles standards pour lequel la fonction d'influence n'est pas sous-modulaire. Nous proposons ensuite, en utilisant des graphes jouets et un réseau social réel, une étude de différents algorithmes de maximisation de l'influence sur ce modèle ainsi que sur le modèle standard IC : des heuristiques simples, la méthode gloutonne, une généralisation de celle-ci et une méthode d'optimisation de fonctions sous-modulaires. Nous montrons que même si la fonction d'influence n'est pas sous-modulaire, l'algorithme glouton obtient de bons résultats tout en gardant une complexité permettant le passage à l'échelle.*

ABSTRACT. *Influence maximization is a NP-hard problem depending on the diffusion of information in social networks. The Greedy hill climbing algorithm have been proved a good approximation if the influence fonction we try to optimize is submodular, which is the case for standard diffusion models.We present a diffusion model not equivalent to standard models for which the influence function is not submodular. Then we propose, using toy graphs and a real social network, a study of different influence maximization algorithms on this model and on the standard model IC: some basic heuristics, the greedy hill climbing method, a generalization of the greedy method and an optimization method for submodular functions. We show that even if the influence function is not submodular, the greedy algorithm obtain good results while being able to scale efficiently.*

MOTS-CLÉS : *Diffusion de l'Information, Maximisation de l'Influence, Réseaux sociaux, Blogs*

KEYWORDS: *Information and content diffusion, Influence Maximization, Social networks, Blogs*

1. Introduction

Les modèles de propagation dans les réseaux sociaux (dans lesquels les utilisateurs se partagent du contenu) cherchent à reproduire la diffusion de l'information entre les utilisateurs. Être capable de bien modéliser cette diffusion permet leur utilisation pour de nombreuses applications, telles que l'identification des nœuds d'influence, le choix des diffuseurs initiaux d'un contenu pour obtenir une propagation maximale, ou encore l'identification des liens à supprimer pour limiter la diffusion (par exemple pour stopper les rumeurs). Nous nous intéressons dans cet article à la détection des nœuds d'influence et plus particulièrement au problème de maximisation de l'influence. Lorsque l'on veut diffuser un contenu dans un réseau social, on cherche souvent à atteindre un maximum de personnes. Pour ce faire, ce problème cherche les diffuseurs initiaux qui entraîneront une propagation maximale. Ce problème étant NP-difficile pour les modèles standards, on recherche des méthodes permettant de l'approximer au mieux.

Nous considérons dans cette étude un graphe social orienté $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ composé d'un ensemble de nœuds/utilisateurs $\mathcal{N} = \{n_1, ..., n_N\}$ et d'un ensemble de liens $\mathcal{E}$. Un utilisateur n_i est lié à un autre utilisateur n_j si $(n_i, n_j) \in \mathcal{E}$. Dans tous les modèles présentés, on s'intéresse aux utilisateurs qui diffusent un contenu. Un utilisateur ayant diffusé un contenu est considéré comme actif (ou contaminé). Un utilisateur est inactif tant qu'il n'a pas diffusé le contenu. Dans ce processus, un utilisateur ne peut pas redevenir inactif s'il est devenu actif : une fois qu'un contenu a été diffusé, c'est définitif. On appelle le contenu diffusé q^k. La suite de cet article s'organise de la façon suivante : la section 2 présente les travaux ayant été effectués dans le domaine de la diffusion de l'information et de la maximisation de l'influence. Nous présentons dans la section 3 le modèle centré utilisateur RUC, prouvé non equivalent au modèle à cascades indépendantes IC. Nous utilisons ces deux modèles dans notre étude. La section 4 donne une définition formelle du problème de maximisation de l'influence et des algorithmes pour l'approximer. Nous montrons dans la section 5 les résultats obtenus et concluons dans la section 6.

2. Travaux reliés

La plupart des modèles de propagation se classent en deux grandes familles : les modèles de contagion et les modèles d'influence. Dans les modèles de contagion, les utilisateurs peuvent être contaminés au moment où ils sont en contact avec un autre utilisateur contaminé. Ces modèles ont été proposés à l'origine en épidémiologie pour modéliser et comprendre la diffusion de maladies au sein d'une population. La famille de modèles SI (*Susceptible, Infected*) et plus particulièrement le modèle SIR (*Susceptible, Infected, Recovered*) modélisent l'état global de la population à l'aide d'equations différentielles déterminant le volume d'utilisateurs passant d'un état à un autre à chaque étape en fonction de l'étape précédente. [TRO 01] donnent une bonne description de ces modèles. Il est aussi possible d'adopter une approche locale en spécifiant les taux d'infection propres à un utilisateur au contact d'un autre utilisateur

contaminé. Dans ce cas là, le modèle SI se rapproche du modèle IC (*Independant Cascade*), comme montré dans [KIM 07]. Le modèle IC [GOL 01] est basé sur le principe suivant : quand un utilisateur n_j (dans notre cas un nœud du réseau social) est contaminé, il a une unique chance de contaminer chacun de ses voisins sortants n_i avec une probabilité P_{ji} qui dépend de n_j et de n_i. Que la propagation ait lieu ou pas, n_j n'essaiera plus par la suite de contaminer n_i. Les paramètres P_{ji} peuvent être appris par maximum de vraisemblance à l'aide de données observées [SAI 08]. Ces deux modèles SI et IC ont été prouvées equivalents à une percolation de lien sur le graphe social [KEM 03, KIM 07].

Les modèles d'influence, aussi appelés modèles de seuil, considèrent qu'un utilisateur est contaminé si le nombre ou la proportion de ses voisins entrants déjà contaminés dépasse un seuil, spécifique à chaque utilisateur. C'est donc cette pression sociale qui détermine si un utilisateur va être contaminé ou pas. Les premières études de ces modèles sont décrites dans [SCH 71] et [GRA 78][1]. Le prototype de cette famille de modèles est le modèle LT qui, dans sa forme première, considère qu'un nœud n_i du réseau social (i.e. un utilisateur) est contaminé si la somme des poids sur ses liens entrants venant de voisins contaminés est supérieure à un seuil qui lui est propre θ_i. Comme pour le modèle IC, on peut montrer [KEM 03] que le modèle LT est equivalent à une percolation de lien sur le graphe social.

Un certain nombre d'études se focalisent sur la maximisation de l'influence pour ces deux familles de modèles. Ce problème a été, à notre connaissance, étudié pour la première fois dans [DOM 01] puis plus tard dans [KEM 03], [KIM 07] et [LES 07]. Il est connu pour être NP-difficile pour les modèles précédemment mentionnés et implique souvent une fonction d'influence à maximiser sous-modulaire qui permet l'utilisation d'un algorithme glouton décrit dans [NEM 78].

3. Modèle de diffusion centré utilisateur

La famille des modèles centrés utilisateur est présenté dans [LAG 13]. Nous n'utilisons dans cette étude que le modèle principal RUC. Ce modèle définit la probabilité qu'un utilisateur devienne actif en prenant en compte trois caractéristiques de celui-ci : son intérêt pour le contenu diffusé, son activité et la pression sociale qu'il subit.

L'intérêt de l'utilisateur pour un contenu est défini par la proximité entre le profil de l'utilisateur et le contenu : $S(n_i, q^k, \theta_s) = sim(p^i, q^k) - \theta_s$. où p^i est le profil de l'utilisateur n_i, $sim(p^i, q^k)$ est une mesure de similarité entre le profil de l'utilisateur et la description du contenu et θ_s est un seuil permettant de définir si les deux objets sont suffisamment proches. En clair, si la similarité est plus grande que le seuil, la proximité sera positive. Elle sera négative dans le cas inverse. L'activité d'un utilisateur $Act(n_i)$ représente le caractère actif ou passif d'un utilisateur. Elle peut être directement calculée à partir des observations faites sur chaque utilisateur (sur un jeu de données d'entraînement). Elle correspond au ratio entre le nombre de

1. Le modèle LT (*Linear Threshold*) est souvent associé au modèle de Granovetter

contenus qu'un utilisateur a vu et repartagé et le nombre de contenus qu'il a vu et n'a pas repartagé. Nous introduisons un seuil θ_w similaire au seuil de similarité θ_s : $W(n_i, \theta_w) = Act(n_i) - \theta_w$. Comme la précédente, cette caractéristique utilisateur sera donc positive si elle doit améliorer la probabilité de l'utilisateur de diffuser, nulle si elle ne doit pas l'influencer et négative sinon. Enfin la pression sociale $SP(n_i, q^k, t)$ représente le fait que la confiance qu'un utilisateur a sur l'intérêt d'une information est croissante avec le nombre de sources différentes. Si un personne nous parle d'un événement, on pourra ne pas y faire attention, alors que si dix personnes nous parlent du même événement on sera forcément au courant et on se dira que ça peut intéresser d'autres gens. Elle correspond donc au nombre de voisins qui ont déjà diffusé le contenu.

En utilisant ces caractéristiques utilisateur, il est possible de définir des fonctions d'aggrégation pour chaque utilisateur, contenu et étape de temps qui serviront a construire les fonction de probabilité de diffusion. Nous avons opté ici pour une simple combinaison linéaire des caractéristiques utilisateur du à leur indépendance :

$$f_\lambda(n_i, q^k, t) = \lambda_0 + \lambda_1 S(n_i, q^k, \theta_s) + \lambda_2 W(n_i, \theta_w) + \lambda_3 SP(n_i, q^k, t) \quad (1)$$

Les paramètres λ_0, λ_1, λ_2 et λ_3 contrôlent l'influence de chaque dimension de la diffusion. Ces paramètres sont globaux (les mêmes pour tous les utilisateurs). Cela veut dire que l'influence de chaque caractéristique par rapport aux autres sera la même pour tous les utilisateurs. Ce n'est pas pour autant que les valeurs de ces caractéristiques seront les mêmes pour deux utilisateurs différents.

3.1. *Modèle probabiliste*

De part les définitions précédentes, la probabilité d'un utilisateur de diffuser une information doit être forte quand $f_\lambda(n_i, q^k, t)$ est grand, c.a.d. quand :

– l'intérêt thématique de l'utilisateur pour le contenu est grand

– l'activité de l'utilisateur est grande

– la pression sociale subit par l'utilisateur est forte

Ces contraintes sont naturellement obtenus en utilisant une fonction logistique qui agit comme une fonction de seuil continu. De plus, un utilisateur ne peut rediffuser un contenu que s'il l'a vu, autrement dit si un de ses voisins entrants a déjà partagé le contenu. La probabilité de diffusion d'un contenu q^k par un utilisateur n_i à l'étape de temps t est donc la suivante :

$$P(n_i, q^k, t) = \begin{cases} (1 + e^{-f_\lambda(n_i, t, q^k)})^{-1} \text{ si } SP(n_i, q^k, t) > 0 \\ \\ 0 \text{ sinon} \end{cases} \quad (2)$$

avec $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ et $\lambda_3 \geq 0$. Si un paramètre vaut 0, la caractéristique utilisateur correspondante n'aura aucun impact sur la probabilité de diffusion. Si le paramètre est

positif, la caractéristique utilisateur correspondante aura un apport positif si elle est positive est négatif si elle est négative.

Le modèle RUC (Reinforced User-Centric model) prend en compte la dimension temporelle pour la diffusion. Un utilisateur n'est pas dans un système binaire dans lequel il est soit actif soit inactif mais on considère sa probabilité d'être actif (et par conséquent sa probabilité d'être inactif). Le temps est représenté de façon discrète, mais un utilisateur peut diffuser un contenu à n'importe quelle étape qui suit sa première prise de connaissance du contenu. Un utilisateur qui voit un contenu pour la première fois chez un de ses voisins à l'étape t aura une probabilité de le rediffuser à l'étape $t+1$, mais aussi à l'étape $t+2$ ainsi que toutes les suivantes, amenant un renforcement dans sa probabilité d'avoir diffusé le contenu. On définit par $P(n_i, q^k, \leq t)$ la probabilité que l'utilisateur n_i ai diffusé le contenu q^k avant l'étape de temps t.

La dynamique du système de diffusion évolue étape après étape, la probabilité d'être actif d'un utilisateur à une étape $t+1$ étant sa probabilité d'avoir été actif à l'étape de temps t à laquelle s'ajoute sa probabilité de s'être activé entre les étapes t et $t+1$:

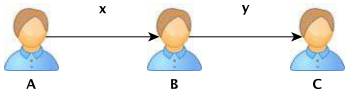
$$P(n_i, q^k, \leq t+1) = \quad P(n_i, q^k, \leq t) + (1 - P(n_i, q^k, \leq t))P(n_i, q^k, t) \qquad (3)$$

Par définition, $P(n_i, q^k, \leq 0) = 1 \Leftrightarrow n_i$ est un diffuseur initial.

3.2. *Non equivalence entre RUC et une percolation de lien*

Comme dit précédemment dans la section 2, les modèles standards de seuil linéaire (LT) et à cascades indépendantes (IC) ont étés prouvés équivalents à une percolation de lien. Le principe d'une percolation de lien sur un graphe est de choisir pour chacun des lien s'il est ouvert ou fermé. Un lien ouvert signifie que la probabilité associée vaut 1, elle vaut 0 pour un lien fermé. A chaque sous-graphe possible du graphe d'origine est associé une probabilité d'apparition.

Theorème 1. *Le modèle RUC défini sur un graphe $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ n'est pas équivalent à un processus de percolation de lien sur $\mathcal{G}$.*



(a) Graphe social

probabilité	x	y
p_1	fermé	fermé
p_2	fermé	ouvert
p_3	ouvert	fermé
p_4	ouvert	ouvert

(b) table des probabilités

Figure 1. *Exemple de percolation de lien*

Démonstration. Afin de prouver le théorème 1, il suffit d'exhiber un graphe sur lequel une instance du modèle RUC ne peut pas être équivalente à une percolation de lien.

La figure 1(a) définit un tel graphe. Après deux étapes de temps, tous les utilisateurs du réseaux auront donc une probabilité non nulle d'avoir reçu le contenu initialement diffusé par A. L'ensemble $\mathcal{V}$ des vecteurs définissant tous les états possible du graphe $\mathcal{G}$ contient quatre vecteurs, et la percolation de lien sur ce graphe est caractérisée par trois probabilités (p_1, p_2, p_3), la quatrième étant définie par : $p_1 + p_2 + p_3 + p_4 = 1$. La figure 1(b) donne la table des probabilités associée.

Si le modèle RUC est équivalent à une percolation de lien, les probabilités que les utilisateurs B et C soient actifs après deux étapes de temps avec le modèle RUC sont les suivantes : $P(B, q^k, \leq 2) = p_3 + p_4$ et $P(C, q^k, \leq 2) = p_4$, ce qui conduit à :

$$\begin{aligned} p_3 &= P(B, q^k, \leq 2) - P(C, q^k, \leq 2) \\ &= P(B, q^k, 0) + (1 - P(B, q^k, 0)) \times P(B, q^k, 1) - P(C, q^k, 1) \end{aligned}$$

Or cette valeur peut être négative si le contenu est plus dans les centres d'intérêts de l'utilisateur C que de l'utilisateur B, ce qui montre que modèle RUC n'est pas équivalent à une percolation de lien, et n'est donc équivalent ni au modèle à cascades indépendantes IC, ni au modèle de seuil linéaire LT. □

4. Maximisation de l'influence

Le problème de maximisation de l'influence (IM) est définit sur un graphe social $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ pour un modèle de diffusion $\mathcal{M}$ et un nombre $\kappa \leq |\mathcal{V}|$. En partant du principe que κ utilisateurs du réseau diffusent une information, le but est de savoir lesquels entraîneront une diffusion maximale. Selon le modèle pour lequel on cherche à maximiser l'influence, nous aurons besoin de paramètres supplémentaires comme les profils des utilisateurs, le contenu diffusé q^k, etc. On définit la fonction σ que l'on cherche à optimiser comme l'influence du contenu diffusé sur le réseau social. En d'autres mots il s'agit du nombre d'utilisateurs actifs à la fin d'une diffusion. Dans la pratique, un certain nombre de modèles étant stochastiques, elle correspond à l'espérance du nombre d'utilisateurs ayant diffusé le contenu : $\sigma_{\mathcal{M}}(e, \mathcal{G}) = E[|C^k(T^k)|]$ où e est un ensemble de κ diffuseurs initiaux, $\mathcal{M}$ un modèle, T^k est le nombre d'étapes de temps à laquelle on stoppe l'observation de la diffusion et $C^k(t)$ est l'ensemble des utilisateurs ayant été contaminé avant l'étape t. Le problème de maximisation de l'influence est défini de la façon suivante :

$$IM_{\mathcal{M}}(\mathcal{G}, \kappa) = argmax_{e \subseteq \mathcal{V}, |e| = \kappa} \quad \sigma_{\mathcal{M}}(e, \mathcal{G}) \tag{4}$$

Il s'agit de l'ensemble de κ utilisateurs qui entraînera une diffusion maximale au sein du réseau.

4.1. *Un problème NP-difficile*

Le problème de maximisation de l'influence a été prouvé NP-difficile pour les modèles standards IC et LT [KEM 03], Il est possible de réduire le problème de cou-

verture d'ensemble, qui est lui même NP-difficile, à une instance du problème de maximisation de l'influence.

On peut effectuer une réduction similaire en utilisant le modèle RUC. Le problème de couverture d'ensemble consiste à trouver, à partir d'un ensemble d'ensembles, une couverture de l'univers en ne choisissant qu'un certain nombre de ces ensembles. De manière plus formelle, on possède un ensemble $\mathcal{C}$ de sous-ensembles de l'univers $\mathcal{U}$ et un nombre κ tel que $\kappa \leq |\mathcal{C}|$. Le problème de couverture d'ensemble cherche à trouver un famille $\mathcal{F}$ d'éléments de $\mathcal{C}$ tel que : $|\mathcal{F}| \leq \kappa$ et $\cup_{f \in \mathcal{F}}(f) = \mathcal{U}$. En d'autres termes, la famille $\mathcal{F}$ doit contenir au maximum κ éléments et couvrir l'univers.

Pour une instance du problème de couverture d'ensemble, on crée une instance du problème de maximisation de l'influence comme suit :

– Un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ avec deux types de nœuds :

- Un nœud pour chaque élément de l'univers : $\forall e \in \mathcal{U}, n_e \in \mathcal{V}$

- Un nœud pour chaque sous-ensemble de l'univers : $\forall c \in \mathcal{C}, n_c \in \mathcal{V}$

– Un lien qui part de chaque ensemble vers chacun des éléments qu'il contient : $\forall c \in \mathcal{C}, \forall e \in c, (n_c, n_e) \in \mathcal{E}$

– On utilise une instance particulière du modèle RUC dans laquelle tous les paramètres λ_i sont fixés à 0 ;

– On cherche un ensemble de κ nœuds qui donnent une diffusion maximale

En résolvant le problème de maximisation de l'influence, on obtient une probabilité d'être actif pour chacun des utilisateurs du réseau. Nous allons retrouver trois classes d'utilisateurs : ceux initiateurs de la diffusion, ceux ayant été atteints par la diffusion, et ceux n'ayant pas été atteints par la diffusion. Une propriété de l'instance du problème telle que nous l'avons définie est que pour tous les utilisateurs d'une même classe leur probabilité d'être actif à chaque étape de temps est la même :

– si n_i est initiateur : $P(n_i, c^k, \leq t) = 1$

– si n_i est atteignable[2] : $P(n_i, c^k, \leq t) = 1 - \left(\frac{1}{2}\right)^t$

– si n_i est non atteignable : $P(n_i, c^k, \leq t) = 0$

Nous rappelons que le graphe ne contient que des chemins de longueur 1 et que donc un utilisateur est soit atteignable dès la première étape de temps, soit ne le sera jamais. La réponse au problème de couverture d'ensemble sera donc oui si $IM_{RUC}(\mathcal{G}, \kappa) \geq |\mathcal{U}| \times (1 - \left(\frac{1}{2}\right)^t) + \kappa$ et non sinon.

4.2. *Algorithmes de maximisation de l'influence*

Ce problème étant NP-difficile, il est légitime de chercher une approximation du résultat optimal. Nous utilisons dans cette étude des heuristiques simples, l'algorithme

2. Somme des premiers éléments d'une suite arithmetico-géométrique

gloutons ayant une garantie sur sa qualité si la fonction à optimiser est sous-modulaire ainsi qu'une variante de celui-ci, et un dernier algorithme de maximisation de fonction sous-modulaire qui n'est, à l'origine, pas créé pour le problème de maximisation de l'influence [KAW 09].

Heuristiques simples

Une manière de chercher à maximiser l'influence est de choisir des propriétés qui semblent diriger l'influence. Nous proposons d'utiliser ici deux heuristiques simples dans lesquelles on choisit les κ nœuds un par un, soit par celui qui a le plus grand degré sortant, soit celui qui est le plus central (qui a la distance moyenne avec les autres nœuds la plus petite).

Algorithme glouton "greedy hill climbing"

Le principe est de dire que si on ne peut pas trouver l'ensemble de κ initiateurs apportant une diffusion maximale, on les choisit un par un. Ainsi, l'algorithme commence par choisir le meilleur initiateur. Ensuite, il choisit le second qui offre le meilleur gain marginal par rapport au premier utilisateur déjà choisi. L'algorithme continue ensuite jusqu'à avoir sélectionné un ensemble de κ utilisateurs. Cet algorithme offre une bonne approximation quand la fonction que l'on veut maximiser respecte certaines propriétés.

Theorème 2. *[NEM 78]. Pour une fonction non négative, monotone*[3] *et sous-modulaire f, soit A un ensemble de κ utilisateurs obtenus par l'algorithme "Greedy Hill Climbing" maximisant la fonction f. Soit A^* l'ensemble qui maximise la valeur de f pour κ éléments. Alors $f(A) \geq (1 - 1/e)f(A^*)$, en d'autres mots, A est une $(1 - 1/e)$-approximation.*

Ce théorème a été utilisé dans beaucoup d'applications d'optimisation discrète [WOL 99].

Il existe une généralisation [DU 08] de cet algorithme qui consiste à ne pas choisir les utilisateurs un par un mais n par n. L'avantage est que théoriquement on obtient une meilleure approximation (voir l'optimal si on choisit $n = \kappa$), l'inconvénient étant que la complexité augmente drastiquement.

4.3. *Toutes les fonctions d'influence ne sont pas sous-modulaires*

Pour le modèle IC, la fonction d'influence est sous-modulaire [KEM 03]. Malheureusement, la borne de l'algorithme de "Greedy Hill Climbing" n'est pas valide pour résoudre la maximisation de l'influence en utilisant le modèle RUC.

Proposition 1. *La fonction σ_{RUC} n'est pas sous-modulaire.*

Démonstration. Nous montrons un contre-exemple en utilisant le graphe social présenté dans la figure 2. Si σ_{RUC} est sous-modulaire, alors

3. Les fonctions de diffusion des modèles étudiés sont non négatives et monotones de par la définition des modèles

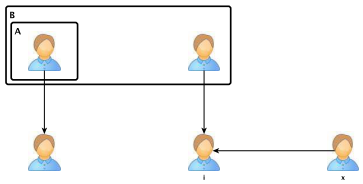


Figure 2. *Graphe social avec 2 ensembles d'utilisateurs A et B tels que $A \subseteq B$*

$$\sigma_{RUC}(A \cup \{x\}) - \sigma_{RUC}(A) \geq \sigma_{RUC}(B \cup \{x\}) - \sigma_{RUC}(B)$$

$$\Leftrightarrow \qquad \frac{1 + e^{-\alpha - 2\lambda_3}}{1 + e^{-\alpha - \lambda_3}} \geq \frac{1}{2}$$

Si $\lambda_3 = 2$ et $\alpha = -4$ alors $\frac{2}{1+e^2} < \frac{1}{2}$. Ce qui amène une contradiction, donc σ_{RUC} n'est pas sous-modulaire. □

5. Expérimentations

Notre but ici est de comparer les choix d'optimisation de différents algorithmes de maximisation de l'influence. Pour ce faire nous avons choisi de comparer :

– l'algorithme glouton "Greedy Hill Climbing", *Greedy-1* ;

– la généralisation de l'algorithme glouton pour $n = 2$, *Greedy-2* ;

– un algorithme de maximisation de fonctions sous-modulaires *SubMax* [KAW 09] ;

– les heuristiques de degré sortant et de centralité

Nos expériences sont effectuées sur deux types de graphes. Afin de comparer les différentes méthodes avec l'optimal, nous générons une centaine de petits graphes (entre 10 et 30 utilisateurs et entre 15 et 180 liens) sur lesquels nous exécutons les différents algorithmes. Le second type de graphes que nous utilisons est un réseau social réel de blogs tirés de Memetracker [LES 09] pour lequel nous n'avons gardé que les 5000 utilisateurs les plus actifs pendant un mois. Nous n'avons pas de connaissances à priori sur les valeurs des paramètres des deux modèles, c'est pourquoi nous les générons aléatoirement.

Le tableau 1 montre la valeur d'influence maximale en fonction du nombre d'initiateurs obtenus sur l'ensemble de graphes jouets pour chacun des deux modèles. On remarque deux phénomènes importantes : tout d'abord, les algorithmes Greedy et SubMax obtiennent de meilleurs résultats que les heuristiques. Les heuristiques définissent les raisons d'une grande diffusion basé par une seule caractéristique indépendante du modèle alors que les algorithmes *Greedy* et *SubMax* utilisent directement les diffusions dictés par le modèle. L'heuristique qui choisi les nœuds par leur degré semble néanmoins être meilleure que celle les choisissant par leur centralité dans le réseau. Le second point est que ces algorithmes sont très proches de l'optimal. *Greedy-2*

	IC					RUC				
$\#initiateurs$	1	2	3	4	5	1	2	3	4	5
$Optimal$	7.80	10.02	11.68	13.17	13.92	7.85	12.70	15.91	17.90	19.15
$Greedy-1$	7.80	9.76	11.17	12.55	13.32	7.84	12.60	15.65	17.54	18.79
$Greedy-2$	7.80	10.02	11.41	12.84	13.58	7.84	12.70	15.72	17.69	18.89
$SubMax$	7.80	9.91	11.55	12.75	13.54	–	–	–	–	–
Degré	6.25	8.41	10.11	11.51	11.75	7.84	11.99	14.43	16.05	17.20
Centralité	6.14	8.12	9.68	11.01	11.35	6.48	10.58	13.20	14.95	16.14

Tableau 1. *Valeur d'influence maximale obtenu sur un ensemble de petit graphes sociaux jouets générés aléatoirement*

$\#initiateurs$	1	2	3	4	5	6	7	8	9	10	11	12
$Greedy-1$	45	79	97	113	126	136	146	153	160	166	172	177
$Greedy-2$	45	79	97	113	126	136	146	153	160	166	172	177
Degré	45	79	97	113	126	136	146	153	158	162	167	171
Centralité	45	79	88	95	102	109	111	113	126	128	134	142

Tableau 2. *Valeur d'influence maximale sur le réseau Memetracker pour IC*

apporte une très légère amélioration par rapport au modèle *Greedy-1* mais pour un coût non négligeable. Nous en parlons dans la section suivante. Quant au modèle *SubMax*, Il obtient des résultats très similaires aux modèles gloutons.

Les tableaux 2 et 3 montrent la valeur d'influence maximale en fonction du nombre d'initiateurs obtenus sur le graphe Memetracker pour chacun des deux modèles. Il n'a pas été possible de calculer les valeurs d'influence pour l'algorithme *SubMax* du à la complexité du problème. En effet, tel que définit, pour calculer l'influence en utilisant le modèle IC il est nécessaire de calculer les probabilités d'activation en parcourant tous les chemins partant des initiateurs. En combinant cette complexité avec celle de l'algorithme, le passage à l'échelle devient compliqué. Comme sur les réseaux jouets, les algorithmes *Greedy* obtiennent de meilleurs résultats que les heuristiques. On remarque cependant que les premiers choix ne diffèrent pas entre tous les algorithmes étudié. Dans un graphe réel on trouve des hubs très importants pour la diffusion qui ont à la fois un degré important et un forte centralité. Ces hubs sont donc choisis à la fois par les heuristiques et par les algorithmes *Greedy*. Il est à noter que la non sous-modularité de la fonction d'influence du modèle RUC ne pose pas de problème quand à la qualité de la méthode *Greedy*. L'heuristique de degré obtient aussi sur ce

$\#initiateurs$	1	2	3	4	5	6	7	8	9	10	11	12
$Greedy-1$	80	124	149	168	181	193	202	211	218	225	232	238
$Greedy-2$	80	124	149	168	181	193	202	211	218	225	232	238
Degré	80	124	149	159	174	186	195	201	206	211	214	216
Centralité	80	124	135	145	154	159	162	163	174	177	183	195

Tableau 3. *Valeur d'influence maximale sur le réseau Memetracker pour RUC*

$Algorithme$	$Optimal$	$Greedy-1$	$Greedy-2$	$SubMax$	$Degre$	$Centralite$
$Reseauxjouets$	$4.05sec$	$0.01sec$	$0.05sec$	$0.43sec$	$0.20ms$	$1.85ms$
$Reseaureel$	$-$	$36.24sec$	$11h$	$-$	$0.32sec$	$0.61sec$

Tableau 4. *Temps d'exécution des algorithmes de maximisation de l'influence pour RUC pour 5 utilisateurs initiaux sur les réseaux jouets et 12 pour le réseau réel.*

graphe de meilleurs résultat que l'heuristique de centralité. On remarque de plus que l'algorithme *Greedy-2* obtient exactement les même résultats que *Greedy-1*. Ceci est du au fait que le graphe réel est localement plus simple que les graphes jouets.

Enfin, la complexité de ces algorithmes est très disparate entraînant une grande différence lors de la comparaison des temps d'exécution.

Le tableau 4 montre les temps de calcul des différents algorithmes pour RUC. On y voit clairement que les deux heuristiques présentées ont un temps de calcul extrêmement court par rapport aux autres méthodes. L'algorithme *Greedy-1* étant plus long à obtenir des résultats, toujours dans un temps acceptable. Les algorithmes *Greedy-2* et *SubMax* n'ont cependant pas la capacité de passer à l'échelle. *Greedy-2* est un niveau de complexité au dessus de *Greedy-1* du au fait qu'il choisit les utilisateurs deux par deux. Enfin, *SubMax* demande encore plus de temps que *Greedy-2* pour donner des résultats et n'est pas utilisable dans le cadre d'un réseau social réel.

6. Conclusion et perspectives

Nous avons présenté un modèle de diffusion de l'information dans les réseaux sociaux d'une classe différente de celle des modèles standards du domaine IC et LT. Néanmoins, pour tous ces modèles, le problème de maximisation de l'influence est NP-difficile. Il est alors cohérent de chercher à approximer la solution optimale. Nous avons étudié dans cet article différents algorithmes cherchant à maximiser la fonction d'influence. Deux paramètres sont à prendre en compte, la qualité de l'approximation, mais aussi la complexité de la méthode. Nous avons vu que la méthode gloutonne "Greedy hill climbing" obtient des résultats très proches de l'optimal sur des réseaux jouets et approxime mieux le problème que des heuristiques simples telles que le choix des initiateurs par leur degré ou leur centralité. De plus, autant la méthode *SubMax* que la généralisation de l'algorithme *Greedy* n'améliorent pas significativement les résultats par rapport à la méthode gloutonne simple mais ont une complexité telle que le passage à l'échelle est compliqué. L'algorithme *Greedy* possède une garanti quant à la qualité de son approximation quand la fonction à optimiser est sous-modulaire mais les résultats sur le modèle RUC, pour lequel ce n'est pas le cas, sont tout aussi bons que pour le modèle IC. Dans un très grand nombre de cas, les algorithmes *Greedy* parviennent à obtenir le résultat optimal. Il serait intéressant d'étudier dans quels cas particuliers (quelles configurations du réseau) ils n'en sont pas capables. Nous avons vu que l'algorithme *Greedy* obtient de bons résultats avec le modèle RUC, un autre travail a venir consiste à essayer de trouver un seuil pour la qualité de cet algorithme quand la fonction d'influence n'est pas sous-modulaire.

7. Bibliographie

[DOM 01] DOMINGOS P., RICHARDSON M., « Mining the network value of customers », *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, ACM, 2001, p. 57-66.

[DU 08] DU D.-Z., GRAHAM R. L., PARDALOS P. M., WAN P.-J., WU W., ZHAO W., « Analysis of greedy approximations with nonsubmodular potential functions », *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '08, 2008.

[GOL 01] GOLDENBERG J., LIBAI B., MULLER E., « Talk of the Network : A Complex Systems Look at the Underlying Process of Word-of-Mouth », *Marketing Letters*, , 2001, p. 211-223, Springer.

[GRA 78] GRANOVETTER M., « Threshold Models of Collective Behavior », *American Journal of Sociology*, vol. 83, n° 6, 1978, p. 1420-1443, The University of Chicago Press.

[KAW 09] KAWAHARA Y., NAGANO K., TSUDA K., BILMES J. A., « Submodularity Cuts and Applications », *NIPS*, 2009, p. 916-924.

[KEM 03] KEMPE D., KLEINBERG J., TARDOS E., « Maximizing the spread of influence through a social network », *KDD '03 : Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, 2003, p. 137-146.

[KIM 07] KIMURA M., SAITO K., NAKANO R., « Extracting influential nodes for information diffusion on a social network », *Proceedings Of The National Conference On Artificial Intelligence*, vol. 22, n° 2, 2007, page 1371, Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999.

[LAG 13] LAGNIER C., DENOYER L., GAUSSIER E., GALLINARI P., « Predicting Information Diffusion in Social Networks using Content and User's Profiles », *IN ECIR*, 2013.

[LES 07] LESKOVEC J., KRAUSE A., GUESTRIN C., FALOUTSOS C., VANBRIESEN J., GLANCE N., « Cost-effective outbreak detection in networks », *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD 07, ACM, 2007, p. 420-429.

[LES 09] LESKOVEC J., BACKSTROM L., KLEINBERG J., « Meme-tracking and the dynamics of the news cycle », *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, KDD '09, ACM, 2009, p. 497-506.

[NEM 78] NEMHAUSER G. L., WOLSEY L. A., FISHER M. L., « An analysis of approximations for maximizing submodular set functions-I », *Mathematical Programming*, vol. 14, n° 1, 1978, p. 265-294, Springer Berlin / Heidelberg.

[SAI 08] SAITO K., NAKANO R., KIMURA M., « Prediction of Information Diffusion Probabilities for Independent Cascade Model », *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III*, KES '08, Springer-Verlag, 2008, p. 67-75.

[SCH 71] SCHELLING T., « Dynamic models of segregation », *Journal of Mathematical Sociology*, vol. 1, 1971.

[TRO 01] TROTTIER H., PHILIPPE P., « Deterministic Modeling Of Infectious Diseases : Theory And Methods », *The Internet Journal of Infectious Diseases*, vol. 1, 2001.

[WOL 99] WOLSEY L. A., NEMHAUSER G. L., *Integer and Combinatorial Optimization*, Wiley-Interscience, 1999.